A Survey of Collaborative Reinforcement Learning: Interactive Methods and Design Patterns

ZHAOXING LI, Durham University, UK
LEI SHI, Durham University, UK
ALEXANDRA I. CRISTEA, Durham University, UK
YUNZHAN ZHOU, Durham University, UK

Recently, methods enabling humans and Artificial Intelligent (AI) agents to collaborate towards improving the efficiency of Reinforcement Learning - also called Collaborative Reinforcement Learning (CRL) - have been receiving increasing attention. In this paper, we provide a long-term, in-depth survey, investigating human-AI collaborative methods based on both interactive reinforcement learning algorithms and human-AI collaborative frameworks, between 2011 and 2020. We elucidate and discuss synergistic analysis methods of both the growth of the field and the state-of-the-art; we suggest novel technical directions and new collaboration design ideas. Specifically, we provide a new *CRL classification taxonomy*, as a systematic modelling tool for selecting and improving new CRL designs. Furthermore, we propose *generic CRL challenges* providing the research community with a guide towards effective implementation of human-AI collaboration. The aim is to empower researchers to develop more efficient and natural human-AI collaborative methods that could utilise the different strengths of humans and AI.

CCS Concepts: • **Computing methodologies** → **Reinforcement learning**; • **Human-centered computing** → **Collaborative interaction**.

Additional Key Words and Phrases: Collaborative RL, Interactive Methods, Design Patterns

ACM Reference Format:
Zhaoxing Li, Lei Shi, Alexandra I. Cristea, and Yunzhan Zhou. 2021. A Survey of Collaborative Reinforcement Learning: Interactive Methods and Design Patterns. In *Designing Interactive Systems Conference 2021 (DIS '21), June 28-July 2, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3461778.3462135

1 INTRODUCTION

With the rapid development of Artificial Intelligence (AI) in recent years, the mainstream media has two opposing views: AI will save the world or destroy it. AI is described as the 'saviour' to free humans from labour, and it is also described as the 'devil' who takes away workers' jobs [78]. Regardless of the viewpoint, it is clear that AI will play an important role in the future world. John Searle proposed three stages of AI development, which include *weak AI*, *strong AI* and *super AI* [58]. Due to the limitations of current technology, Searle believes that we are supposed to be in the weak AI stage for a long time. That is, at this current stage, AI often performs much worse than humans in highly complex decision-making tasks that require consideration of morality and risk, but much better in tasks with well-specified feedback and large scale data. Therefore, the views of popular media are still too far away from our real

Manuscript submitted to ACM

world. Exploring thus how to make humans and AI cooperate, to complement each other's shortcomings, is the best way forward for the immediate future.

Due to its strong potential and firm theoretical foundation, Reinforcement Learning (RL) has recently been regarded as one of the most attractive research fields in AI technologies [61]. Compared with other AI technologies, reinforcement learning is the most widely used algorithm in decision-making tasks [61]. Reinforcement learning mainly emphasises the interactions between the agent and the environment. This is different from supervised learning and unsupervised learning. In supervised learning, the algorithms learn the mapping relationship between input x and label y through a set of corresponding data; and in unsupervised learning, algorithms learn features or patterns from unlabeled data for categorisation [7]. In contrast, reinforcement learning is mainly manifested via "teacher signals". The reinforcement signal provided by the environment is an assessment of the quality of the generated action by the agent (usually a scalar signal), instead of telling the agent how to produce the correct action [56]. Its goal is to obtain information from these interactions and learn the relationship between states and actions. The mapping guides the agent to make optimal decisions based on these states towards maximising cumulative rewards [61]. From 1995, when Tesauro proposed TD-Gammon to play backgammon [63], to 2016, when AlphaGo defeated a professional human GO Player, reinforcement learning has made great progress [19]. Now, it has gradually been applied in fields such as games [37], robotics [35], computer vision [7], natural language processing [8], and recommendation systems [54]. For example, the Open AI[1] team proposed an interactive reinforcement learning method that uses human feedback to learn summarisation [59]. Another very recent project proposed by this team, GPT-3, has also made great achievements in the NLP field [8]. However, reinforcement learning also faces many challenges. On the one hand, the agent needs a lot of prior knowledge to understand its state in a complex environment. On the other hand, even if the agent has been given well-specified feedback, due to the inexplainability and incomprehensibility caused by the black-box nature of neural networks, it is still insufficient for the agent to decide on the precise next action.

Therefore, at the moment, how humans and machines can cooperate and complement each other's shortcomings becomes particularly important. In this survey, we summarise the current interactive methods and put forward our own views and suggestions. We reviewed classic human-machine interaction techniques in the field of engineering. These techniques have had profound impacts on the development of the Human-Computer Interaction (HCI) field. We hope these techniques could inspire other researchers to develop novel human-AI collaborative methods in a more consistent manner.

Showcasing the importance of the field, some recent good survey papers on Collaborative Reinforcement Learning (Collaborative RL, or CRL) have appeared. These literature reviews have covered wide topics, such as CRL in general [2] or specific subjects within the CRL field, such as safe RL [18], inverse RL [17], the explainable RL [52]. Other literature reviews have focused on details of design methods, such as user feedback and testbeds of the environment [38].

However, none of these studies investigates the Human-AI collaboration from the engineering perspective. Furthermore, these studies do not consider proposing potential future directions for the development of human-AI collaboration, especially from the engineering perspective. Thus, *our survey is focusing on predicting the pattern of human-AI interaction from the engineering perspective*, which could provide implementers with a design method that combines archetypes in a macro-view and specific tools in a micro-view [11]. Moreover, we provided a new classification method, as a systematic modelling tool, which may provide researchers with the next stage research directions of human-AI collaborative designs.

[1]The website of OpenAI: www.openai.com

Therefore, the main contributions of this survey lie in the following aspects:

(1) First, we take stock and summarise the latest collaborative reinforcement learning algorithms, defining the state-of-the-art at the start of this new decade.
(2) Second, we summarise the most important *human-machine collaborative patterns from an engineering perspective*, which will provide better guidance for Human-Computer Interaction (HCI) researchers and practitioners.
(3) Third, we have provided a *new CRL classification* and *CRL taxonomy* as a systematic modelling tool to help researchers select and improve new CRL designs.
(4) Fourth, we propose concrete *generic CRL challenges* for future research in this area, as a guild towards effective human-AI collaboration.

2 METHODOLOGY AND SCOPE

In recent years, collaborative or interactive reinforcement learning has become an emerging branch of machine learning. When performing a brief search on Google Scholar with the keywords 'interactive AI', 'cooperation', 'reinforcement learning', and 'HCI' (Human-Computer Interaction) for the period from 2018 to 2021, we found that there are many surveys or literature reviews published. For example, only between January 2020 and January 2021, Najar and Anis published a literature review of reinforcement learning based on human advice [51]. Gomez and Randy published a review of human-centered reinforcement learning [41]. Puiutta and Erick presented a literature review of explainable reinforcement learning [52]. Arzate and Chirstian presented a survey on the design principles and open challenges of interactive reinforcement learning [6]. Suran and Shweta presented a review on collective intelligence [60]. It can be seen that this research direction is receiving more and more attention from the community. We consider that these literature surveys have direct relevance for our work, as they seem to indicate that interactive reinforcement learning and explainable reinforcement learning can potentially be applied to collaborative reinforcement learning.

Therefore, instead, we focus here on the unexplored area of works that have been successfully applied to collaboration between human and machines (AI agents). We further refine our data pool, specifically narrowing the selection to the following selected target areas: HCI, Human-AI Collaboration, Reinforcement Learning, Explainable AI, published over the recent decade - the time period between 2011 and 2020. In total, our search resulted in 116 articles, which contain keywords including *collaborative reinforcement learning*, *interactive reinforcement learning*, *human-computer interaction*, and *engineering design patterns*. They were published in journals and conferences, including top venues such as AAAI[2], CHI[3], UbiComp[4], UIST[5], IEEE[6]. In the next step, we excluded 47 articles deemed, after manual inspection, of lower relevance, leaving 69 articles serving as the source of this survey. Although the interaction between humans and AI (agents) is an emerging research direction, the research on the interaction between humans and computers can be traced back long ago. Many patterns of human-computer interaction have been proposed by the community. For example, in 1983, Hollnagel and Woods proposed the Cognitive Systems Engineering (CES) model [24]. Schmidt *et al.* introduced a conceptual framework in 1991 that divided human-computer collaboration into an augmentative level, an integrative level and a debative level [55]. Johnson *et al.* proposed a co-active design pattern in human-AI joint activity in 2009 [29]. Among the above approaches, the framework proposed by Schmidt [55] is the most widely

[2]The website of AAAI: www.aaai.org/
[3]The website of CHI: chi2021.acm.org/
[4]The website of UbiComp: www.ubicomp.org/
[5]The website of UIST: uist.acm.org/
[6]The website of IEEE:www.ieee.org/

Table 1. *CRL Classification* applied to the Pool of Papers about Collaborative RL, published between 2011-2020

	Interactive Methods	References
Interactive methods	Explicit Methods	[34], [45], [46], [66],[65],[34],[73], [12]
	Implicit Methods	[71], [50], [40], [16], [3], [20], [21], [47],[36], [75]
	Multi-model Methods	[53], [74],[27], [39],[23]
Algorithmic models	Reward-based methods	[67], [34], [4],[3], [20], [21], [47]
	Policy-based Methods	[22], [46], [5], [20], [21]
	Value Function based methods	[62], [42], [49], [57], [50]
	Exploration-process methods	[65], [64], [47], [36], [75], [28], [55]
Design patterns	Augmentative Level Cooperation	[50], [1], [77], [70],[44],[48]
	Integrative Level Cooperation	[28], [79], [69], [10], [9], [43], [55]
	Debative Level Cooperation	[55], [32], [26], [18], [15], [6], [5]

used in engineering and computer science. In this survey, we summarise the main interactive methods of cooperative reinforcement learning.

In previous work, the article by Najar and Anis mainly targeted physical interaction between humans and AI from a human perspective [51]. In the work of Arzate and Chirstian, more attention was paid to the algorithmic model of the AI agent [6]. After reviewing the literature of human-machine interaction in traditional engineering, we found that the interaction between humans and AI also conforms to these patterns. In particular, Schmidt's model not only combines the interactive methods and algorithmic models, but it also provides different design ideas according to different human-AI collaborative levels [55]. Therefore, we used an inductive method to organise the literature we collected, and proposed our new classification method inspired by the traditional engineering human-machine interaction research.

As a human-AI collaborative model usually involves two roles (human and AI) [31] and the collaboration patterns between them, we divided the articles collected into three categories (see Figure 1). From a human perspective, we focus on how humans interact with AI agents. This part is described in the Interactive Methods section. From an AI agent perspective, we focus on how agents accept human instructions or suggestions in algorithm implementation. This part is described in the section 4. From the perspective of collaboration patterns, we focus on how humans and AI can collaborate. This part is described in the section 6. We define these three axes and populate them with representative works from the literature, allowing us to discover, in a structured way, future development trends for CRL (see Table 1).

This taxonomy could be a systematic modelling tool for HCI researchers and designers from the DIS community to select and improve their new CRL designs. AI system researchers and designers have two main tasks in recent years: developing new interactive technologies and applying these technologies to specific scenarios [6]. When researchers start designing an AI system, they could first select a collaborative model from a macro perspective in the *Design Patterns* category. Then, they could select suitable interactive methods and algorithmic models for the specific task requirement categories of the *Interactive Methods* and *Algorithmic Models*.

3 REINFORCEMENT LEARNING

Reinforcement learning is developed from theories of animal learning and parameter disturbance adaptive control. Its basic principle is: if a specific behaviour strategy of the agent leads to positive rewards (reinforcing signals) in the environment, then the tendency of the agent to produce this kind of behaviour strategy will be intensified. The goal of the agent is to find the optimal strategy for each discrete state, to maximise the expected discount rewards [61].

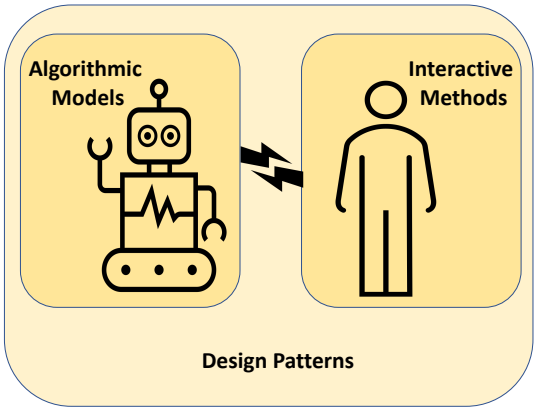

Fig. 1. Human-AI Cooperation Design Model

Reinforcement learning is different from supervised learning, which is mainly manifested via "teacher signals". In reinforcement learning, the reinforcement signal provided by the environment is an assessment of the quality of the generated action by the agent (usually a scalar signal), instead of telling the agent how exactly to produce a correct action. Since the external environment provides very little information, the agent must learn from its own experience. This way, the agent gains knowledge in the environment by "one-by-one evaluation" of actions and improves the action plan to achieve the optimal result. For example, when training a supervised learning model, a set of corresponding data is required, including the input value x and the correct output value y [30]. However, reinforcement learning does not require such a set of pairwise data to train the model. If the reward signal r and the action A were known, the supervised learning algorithm could be used directly. It is not easy to enumerate all actions and corresponding reward signals to train the model in real-world scenarios. Therefore, reinforcement learning is of great significance. In a scene with limited discrete action space, such as a game of Go or Atari, reinforcement learning usually outperforms supervised learning [1].

However, the development of reinforcement learning has encountered obstacles: current reinforcement learning only works well when the environment is definite, i.e., the state of the environment is fully observable. Specifically, in games such as Go, there are clear rules, and the action space is discrete and limited. Besides, most of the applications of reinforcement learning so far are only for playing games such as chess and Atari. It is considered as an effective way to use human's prior knowledge to help reinforcement learning agents improve efficiency and expand usage scenarios. This process requires efficient interactive methods.

Figure 2 presents the most commonly used and highly cited methods and patterns in the Collaborative RL research. The first part collects interactive methods. It includes explicit and implicit interaction methods, as well as multi-module interaction modes [6]. The second part reflects algorithms and models. It contains reward-based methods [4], value-based methods [62], policy-based methods [22], and exploration-process-based methods [65]. The third part are design patterns. These include cognitive systems engineering (CSE) [24], Bosch's framework [68], the Coactive design pattern [28] and Schmidt's framework [79]. This taxonomy could be a systematic modelling tool for researchers to select and

improve their new CRL designs. They could choose a archetype in design patterns for the overall architecture and select suitable interactive methods and algorithmic models for the specific task requirement.

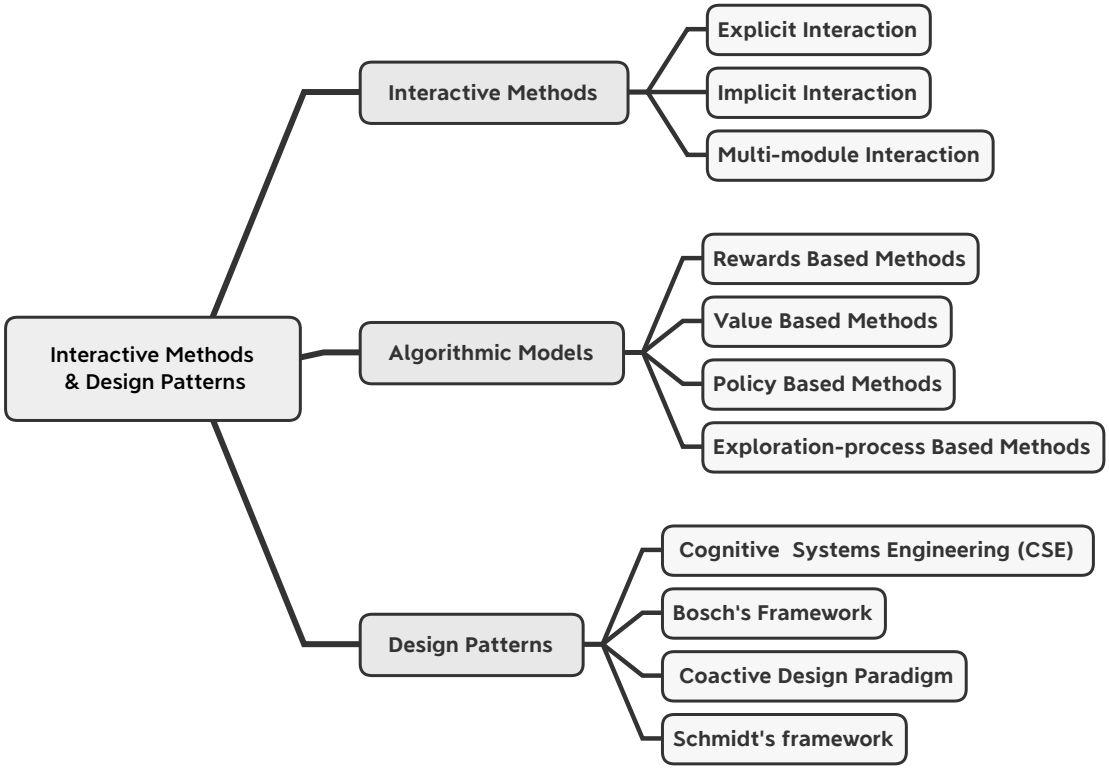

Fig. 2. A new *CRL taxonomy* for interactive methods and design patterns

4 INTERACTIVE METHODS

Traditional reinforcement learning methods require excessive training-time in complex environments, and their applications are often limited to scenarios with clear rules. An effective way to mitigate these limitations is utilising different strengths of human and AI and complementing each other's shortcomings. This approach is coined as Collaborative Reinforcement Learning (CRL). CRL involves human-in-the-loop training to improve the performance of the algorithms or help humans improve decision-making efficiency [6]. Recent research into CRL has focused on developing AI that can communicate with humans in a more natural way [6]. Interactive methods can be divided into explicit and implicit methods. In an explicit method, humans provide the AI agents with clear numerical feedback, explicitly. This method is better for AI agents, which can process the feedback more easily, but it is likely to cause human fatigue and thus lead to inefficiency in a long-term training process. In an implicit method, humans give feedback to AI agents through natural interactions such as posture and gaze, as opposed to clear numerical feedback in explicit methods. This method makes higher demands from the AI agent, but it could improve the fatigue resistance of human

trainers, thereby achieving long-term and stable cooperation [27]. Based on these considerations, in this section, we present the human-AI interactions from the perspective of interactive methods.

4.1 Explicit Interactive Methods

Currently, most AI agents learn from the feedback delivered by humans using explicit interactive methods. Humans deliver the feedback directly via keyboard, slider bar, or mouse, to give the agent clear alpha-numerical feedback [34], [45], [46], [66]. For example, Thomaz and Breazeal proposed a method of sending feedback to the agent by clicking on the sliding bar with the mouse [65]. Knox and Stone proposed the TAMER framework, where an agent can learn from MDP and human advice, via a human trainer clicking the mouse to choose from the actions expected [34]. These methods are more efficient than traditional reinforcement learning, and can also accomplish specific goals in complex environments with human's help.

However, the reaction time of human trainers may cause delayed feedback, so that the agent may not be sure which actions the human feedback was aimed at, especially for agents with frequent actions. A common solution is to set a delay parameter to express the past time-steps. Warnell *et al.* proposed a method to obtain the delay distributions of the human trainers to improve algorithm efficiency [73]. Another method is to use a probability density function to estimate the delay, which was proposed by Knox and Stone [33]. Moreover, these methods may be very unfriendly to non-professional human trainers, who need to spend a lot of time to learn the user interface and the meaning of feedback represented by each operation. At the same time, this kind of interactions may easily make the human trainers feel impatient.

Human trainers can also provide explicit feedback to the agents via hardware delivery methods [12], which could be generally converted into a numeric value directly by the hardware devices, such as keyboards. However, a user-friendlier way is when the agent learns the implicit feedback from natural interactions with trainers.

4.2 Implicit Interactive Methods

Besides receiving feedback via explicit interactive methods from human trainers directly, the agents can also learn via implicit interactive methods.

Implicit interaction methods reduce the learning cost of human trainers who can directly participate in training the agents without special learning. At the same time, a more natural interaction way makes human trainers less fatigued. There are many implicit interactive methods that have been proposed recently. For example, feedback can be based on natural language, facial expressions, emotions, gestures and actions, and the integration of multiple natural interactive methods. Ideally, humans could train the agent as naturally as interacting with humans in the real world. Below, we summarise some of the most popular implicit interactive methods.

Gestural Feedback. Gestures are sometimes considered to be a human unconscious communication. It is also considered to be an effective way that can complement other communication forms, and it is even more useful than other communication methods for users who are speech- or hearing-impaired. For example, Voyles and Khosla proposed a framework which can train robots by imitating human gesture [71]. Moon *et al.* introduced a method of using gestures to command the agent to learn to control a wheelchair [50]. These methods are very friendly to human trainers and do not require special training from them.

Facial Feedback. Li *et al.* trained a mapping model that can map implicit emotions to different explicit feedback data. Facial expressions were marked with different types of feedback in advance, such as 1 for "happy", 0 or -1 for

"sadness" [40]. Based on this work, Gadanho introduced a facial feedback reinforcement learning method, which is based on an emotion recognition system. The system can learn to decide when to change or reinforce its behaviour with Q-learning by identifying human emotions [16]. Arakawa *et al.* introduced the DQN-TAMER model, where an agent can obtain facial expressions from a camera, and then use the facial expression data to map different emotions as implicit rewards to improve the learning efficiency [3]. Veeriah *et al.* proposed a method where the agent can analyse human facial features from photos obtained by the camera to get additional rewards. This way, the agent can quickly adapt to the user's face changes to complete the task [20]. One of the limitations of this kind of methods is that human facial expressions can represent less meaning, and there may be a delay in the conversion of machine recognition expressions into feedback.

Natural Language Feedback. Compared with facial expression and gestural tracking feedback methods, the feedback based on natural language makes it easier to convert the token vector of the sentence into quantitative feedback. Natural language feedback can be transformed and applied in different parts of reinforcement learning, such as rewards, values and policies. Goyal *et al.* introduced the LEARN (LanguagE-Action Reward Network) method, which is a reward shaping method [21]. In the state-action space of the task, if most of the reward signals are 0s, we call it the sparsity of rewards. Sparse rewards may cause the algorithm to converge slowly. Agents need to interact with the environment several times and learn from a large number of samples to reach an optimal solution. One way to solve this problem is to give the agent an additional reward beyond the reward function when the agent takes a right step toward the goal. This process is called reward shaping. Maclin and Shavlik proposed RATLE (Reinforcement and Advice, Consulting Learning Environment) [47]. In this framework, the agent can translate human natural language suggestions into feedback for the Q-value function to accelerate the learning process. Kuhlmann proposed a method based on transforming natural language suggestions into an algorithm-understandable formal language to optimise the learning policy [36]. In addition to the above methods of transforming into different parts of the algorithm, natural language can also directly guide the agent's policy of learning. For example, Williams *et al.* proposed an object-oriented Markov Decision Process (MDP) framework which can map the natural language to rewards feedback [75].

4.3 Multi-modal Feedback

The research above is focused on a single input interaction method. However, in daily human-human interactions, multi-modal interactions are more common and efficient. Multi-mode communication has the following advantages. First, when a single-mode piece of information is disturbed by noise or occlusion, other modes could be used as information supplements. Second, when multi-modal interaction is available, it could improve the robustness and reliability of communication. Quek *et al.* introduced a framework for analysing the mutual support of language and accompanying gestures [53]. Cruz *et al.* proposed a dynamic multi-modal audiovisual interaction framework to allow humans to provide feedback via voices and gestures [74]. Griffith *et al.* [22] introduced a multi-modal interaction method based on hand gesture and speech recognition system, which was limited to operating geometric objects on maps. Weber *et al.* [74] developed a dynamic audiovisual integration method to allow humans to provide information via natural language and gestures. In the above experiments, multi-mode interactions generally performed better than single-mode interactions. Most of the current multi-mode interactions only stay in the combination of two modes, such as any two of voice, gesture, sound and vision. One of the problems of the above multi-modal methods is that they cannot combine different human feedback well. How to find a better way for humans to directly interact with agents

using multiple methods at the same time still needs a lot of explorations. These multi-mode interactive methods can be combined in more forms in the future to develop suitable human-AI cooperation for more scenarios.

Some studies take into account the impact of human fatigue caused by increased training time on the quantity and quality of feedback. As the training time increases, human trainers feel fatigued, thereby reducing the amount of feedback and so the quality of feedback may also decrease [27], [23]. Methods for motivating human trainers to increase interaction enthusiasm with gamification were proposed; such methods have shown to be able to alleviate human trainers' fatigue and improve their efficiency effectively [39].

5 ALGORITHMIC MODELS

In the previous section, we analysed how humans deliver feedback to agents. In this section, we classify algorithmic models based on how agents receive human's feedback.

Reward-based Methods. Reward-based methods accelerate the learning process by adjusting the reward that the agent receives from the environment. Concretely, after the agent receives feedback from the environment, humans could scale up or down the rewards based on their knowledge, which can accelerate the learning process [4]. Thomaz and Breazeal proposed a method for non-expert human trainers. These trainers were able to give a positive or a negative numerical reward to the agent to modify its next action. If the agent received a negative feedback, it would try to withdraw the last action to get a better score [67]. Knox and Stone first introduced the TAMER algorithm, which uses human demonstration as input to guide the agent to perform better [34]. Based on the TAMER method, Riku *et al.* introduced a framework which combines the deep learning method and TAMER, named DQN-TAMER, where rewards were shaped by the human's numerical binary feedback and environments [4].

Reward-based Methods can effectively accelerate the learning process in an environment with sparse rewards, but there are also problems, as follows. The first problem is "credit allocation", particularly in a rapidly changing environment, where humans may be too slow to provide feedback in time. Therefore, how to map human rewards to corresponding actions remains the limiting factor of this method. The second problem is "reward hacking", where the agent may achieve the greatest rewards through methods that are not expected by humans [6].

Policy-based Methods. Policy-based methods modify the learning policy of the agent action process to encourage the action to fit what the human trainers expect [6]. At present, the method that uses human critique for state and action pairs as input to shape agent policy is widely accepted. Griffith *et al.* proposed an optimal policy method based on human feedback, a Bayesian method that takes critiques for each state and action pair as input [22]. MacGlashan *et al.* proposed a Convergent Actor-Critic method, COACH (COrrective Advice Communicated by Humans). This framework allows non-experts to use numerical binary feedback to formulate policies through corrective suggestions [46]. Dilip *et al.* proposed a deep COACH method based on the original COACH, which uses raw pixels as input to train the agent's policy. The authors argued that the use of highly representative inputs facilitates the application of the algorithm in more complex environments [5].

Compared with reward-based methods, the advantage of policy-based methods is that they do not require specific feedback from humans to agents. Nevertheless, humans need to know which strategy may be the best to help the agent. This may have higher requirements for the prior knowledge of human trainers.

Value Function based Methods. The value function based method is to estimate future rewards to get the largest reward possible at the end of the task by using human knowledge [6]. Value function based methods combine the value

representing human preference with the value obtained by the agent from the environment to promote the learning process. Matthew *et al.* proposed a method that combines human preference and agent value called Human-Agent Transfer (HAT) [62]. The algorithm generates a strategy from recorded human trainer preferences, and then uses this strategy to shape the Q-value function. This shaping process provides a stable reward for the state-action pair in the Q-learning process. Brys *et al.* proposed a method that uses human demonstrations as input for a value named RLfD. This method generates a Gaussian function by human demonstration to guide the exploration process of the Q(λ) algorithm [42].

At present, though the value function method is likely to be an effective method to minimise human feedback, there are still only very few studies based on the it.

Exploration Process based Methods. Reinforcement learning is a method in which an agent needs to continuously interact with the environment and complete tasks based on rewards. This means that the agent needs to perform actions that it has not tried before. This process is called an exploration process. In exploration process based methods, humans can reduce agent errors and unnecessary attempts based on their knowledge to improve efficiency [4]. Exploration process based methods aim to minimise the action space by injecting prior human knowledge to guide the agent's exploration to improve learning efficiency. Thomaz and Breazeal conducted an experiment in the game Sophie's Kitchen to evaluate human guidance that helps the agent minimise its action space to accelerate learning efficiency [65]. The results show that human prior knowledge can effectively help the agent reduce low utility attempts, which is more efficient than using scalar reward functions [64]. These methods are considered effective, but they generally need to be trained by humans and this training process requires a lot of professional knowledge and participation.

In general, collaborative reinforcement learning has shown great potential in improving the efficiency of decision-making tasks. However, how to build the environment models, in which humans interact with agents, is a strong limitation. These models should not only consider the effectiveness and efficiency of the interactive methods, but also consider the interpretability, accountability, and possible ethical issues in decision-making. Therefore, in the following sections, we refer to the literature on the pattern of human-machine relations in the engineering field, and propose guidance for the future development of collaborative reinforcement learning methods.

6 HUMAN-AI COOPERATION DESIGN PATTERN

Human-AI cooperation design pattern provides an efficient and reusable solution to designing human-AI collaborative systems [14]. Reliable design patterns could improve the quality, reusability and maintainability of these systems. In this article, we collected the most recognised design patterns in the literature of human-machine cooperation for researchers and designers. We hope it could inspire the design of human-AI cooperative systems.

Compared with human-AI cooperation, human-machine cooperation has long been a concern of researchers. In the early stages of the development of human-machine interactions, domain experts believed that the cooperation between humans and machines was a physical, lower-level type of cooperation [55]. That is, humans only used tools through physical contact, without feedback from machines to humans, which was a kind of unidirectional interaction. After decades of development, there are several recognised human-machine cooperation design patterns, which we summarise as follows:

6.1 CSE Pattern

Hollnagel and Woods proposed Cognitive Systems Engineering (CSE), which operates at the level of cognitive functions [24]. CES introduces a new cooperation framework for human-machine cooperation, where machines use knowledge or the information provided by humans in planning and exploring. This engineering method proposes that the cooperation between humans and machines is at a conscious level of communication. The first mode is the perception mode, which means that the machine is used as a sensory extension to interfere with human activities.

The main challenge at this level is to find appropriate interactive methods to optimise human information processing. CSE was the first framework that was proposed to consider the information exchange relationship between humans and machines. However, it is limited to only considering simple and low-level communications, and does not consider the complex problems and environments.

6.2 Bosch and Bronkhorst's Pattern

Bosch and Bronkhorst defined three levels of Human-AI Cooperation: 1) unidirectional interaction, i.e. either humans help machines or machines provide explanations to humans; 2) bi-directional interaction, and 3) collaboration between humans and machines [68]. The vast majority of the currently existing methods have been focused on only the first level.

The contribution of this framework is to consider the directions of the collaboration between humans and machines; that is, which role is the subject of the task, and which role assists the other subject. It is conducive to build more efficient communication methods. For example, if it is a human-centred framework, more considerations should be given in terms of how to transform 'machine language' such that humans can better understand; if it is a machine-centred framework, more considerations should be given in terms of what knowledge of humans could improve the efficiency of machines.

6.3 Coactive Design Pattern

Johnson*et al.* proposed a Coactive design pattern in human-AI joint activities. In the Coactive Design approach, they experimented with a cooperation system following the perspective of observability, predictability and directability [28]. Observability concerns the ability of both robots (or AI agents) and humans to observe each other's pertinent aspects of status, as well as the knowledge of the team, tasks and the environment. Predictability means that the actions of both robots (or AI agents) and humans can be predicted such that they can rely on each other's actions to perform their own actions. Directability refers to the ability of both robots (or AI agents) and humans to direct each other's behaviour in a complementary manner.

This framework is similar to the Bosch and Bronkhorst's framework, taking into account the direction of interactions and dividing it into different levels. However, it is limited due to the lack of robustness and security considerations.

6.4 Schmidt' s Pattern

Schmidt believes that cooperation should be oriented to different needs, serve different functions, and choose different forms according to different requirements. Cooperation can be summarised into the following: 1) the augmentative level: one role in the partnership (human or AI agent) helps the other perform tasks; 2) the integrative level: both sides of the team share information and assist each other to complete tasks together; and 3) the debative level: tasks are completed through debate and negotiation between humans and AI agents especially when facing complex issues [79].

Not only does this framework consider the information exchange direction in the collaboration, but it also considers different levels of cooperation, as well as the robustness, security and possible ethical issues.

7 DIFFERENT LEVELS OF COLLABORATION

The above patterns frame the modes of human-AI cooperation from different perspectives. At the same time, they are very similar in terms of compartmentalising the cooperation modes or methods, i.e. into single-direction assistance, bi-directional cooperation and higher-stage fused cooperation. In this survey, we adopt a fusion viewpoint that combines interactive methods and design patterns based on Schmidt's cooperation pattern to classify the current collaborative reinforcement learning techniques. Schmidt's model is divided into three levels: augmentative level, integrative level and debative level. We drew a pyramid model based on Schmidt's model (see Figure 3). In the following sub-sections, we list the methods that have emerged at each level and the issues that should be considered. We also discuss the characteristics, advantages and disadvantages of these different methods, and how to develop new methods in the future.

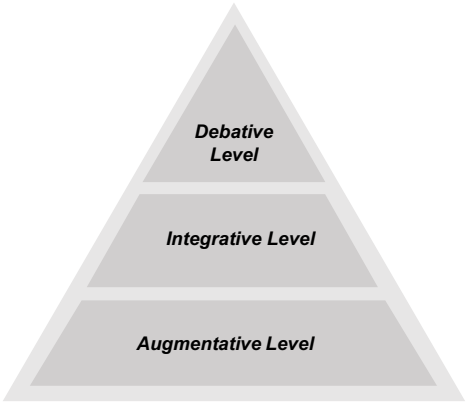

Fig. 3. Different levels of collaboration

7.1 Augmentative Level Cooperation

Cooperation at the augmentative level means that one partner is to compensate for the other's limitations [55]. AI has shown great potential in large-scale data processing with clear rules and has made great progress in fields such as digital image recognition [7], natural language processing [72], among others. Nevertheless, in a complex ambiguous environment, the performance of AI is still far behind humans. The methods proposed by the community in the augmentative level are mainly divided into two types. First, AI plays the main role of decision-making, and human assists AI to improve computing efficiency. In this case, humans use prior knowledge to help the agents specify the state space and get rewards efficiently from the complex environment. Secondly, humans play the main role in decision-making, and AI assists humans in decision making. In this case, the agents provide explanations or assistance to help a human make decisions faster in a simple environment. At the sub-level of humans helping AI agents improve efficiency,

we will classify how they communicate based on which part of the algorithm humans' help can be injected. At the sub-level of AI agents helping humans, we mainly focus on how AI agents can inform humans why they make decisions in a certain way.

7.1.1 Human->AI. In the task of humans assisting AI agents in decision-making, how to efficiently deliver information to agents while reducing human fatigue is the most important consideration. At present, many Human-AI collaborative reinforcement learning algorithms have been proposed. Here, we categorise them into explicit interaction modal, implicit interaction modal and multi-modal methods based on different forms of interactions (detailed description in Section 4). How to find a better way for humans to directly interact with AI agents still needs a lot of exploration.

7.1.2 AI->Human. In the task of AI agents assisting humans in decision-making, the most challenging problem is related to interpretability. Interpretability or explainability refers to the degree to which humans can understand the reasons for a decision-making [25]. The interpretability of AI models refers to the clarification of the internal mechanism and the understanding of the results. The more interpretable the model, the easier it is for people to trust it [52]. Its importance is reflected in the following aspects: in the modelling phase, interpretability can assist developers in understanding the process, making comparisons to select different algorithms, optimising the procedure, and fine-tuning the model; in the operation phase, AI agents can explain the internal mechanism and interpret the model results to the decision-maker (humans). Take a decision-making recommendation model for example, before the model runs, multiple interpretable algorithms with their respective advantages can be provided to humans to choose; and after the model runs, the model needs to explain why it recommended a specific solution for this human.

Patterns underlying the above problems fall into the so-called eXplainable AI (XAI) field, which is widely acknowledged as crucial for the practical deployment of AI models. XAI is an initiative originated by DARPA in 2016 [1]. The main goal was the production of machine learning models, which, using appropriate explanation techniques, will enable humans to understand better and ultimately trust the model's process and results. Two types of explainability are generally proposed in the literature: 1) transparent models, that are embedded inside the operation of the AI algorithms leading to explainability by design, applied to simpler AI algorithms with less accurate results; and 2) post-hoc models, performed after AI algorithms run. This kind of methods is usually more efficient, but its reliability is lower than transparent models. [77].

At present, there are a few intrinsic interpretability RL methods. Verma *et al.* introduced a Programmatically Interpretable Reinforcement Learning method (PIRL) [70]. This method is an upgrade of traditional deep reinforcement learning (DRL). In deep reinforcement learning, due to the black-box nature of neural networks, it is difficult to represent policies. In the PIRL, an advanced human-readable programming language is introduced, to represent neural network policies. Shu *et al.* introduced a hierarchical and interpretable multi-task reinforcement learning framework. In this framework, a complex task is divided into different sub-tasks, and then a hierarchical strategy is used to complete the learning under the 'weak supervision' of humans. This method builds intrinsic interpretability by explaining the relationship between different hierarchies of sub-tasks.

Compared with intrinsic interpretability RL methods, post-hoc methods are simpler in algorithm structure and more efficient in the computing process. At present, many post-hoc methods have been proposed. For example, Liu *et al.* proposed an explainable deep reinforcement learning method based on linear model U-trees [44]. This is a stochastic gradient descent framework that explains complex models by using linear model U-trees to fit Q-functions. There is also a Soft Decision Tree (SDT) method, which provides post-hoc explanations by extracting policies. Madumal *et al.* introduced an explainable method through a causal lens. In this framework, an agent learns to play StarCraft II, a large

dynamic space strategy game [48]. In order to generate an explanation, they simplified the entire game states to 4 basic actions and 9 basic states, and then used these basic causal factors to construct an explanation as to why the agent chooses action A instead of action B.

7.2 Integrative Level Cooperation

Integrative cooperation means integrating different advantages of both parties to complete a task. At this level, humans and agents are considered to be in an interdependent relationship. The main task is divided into several sub-tasks, and humans and agents can choose what they are good at to complete [55]. At the integrative level, the roles of human and AI are equal in the system. The information exchange at this level is generally referred to as 'communication' in the literature [55].

In the following sub-sections, firstly, we summarise the communication methods in this cooperative pattern. Then, we discuss how to make the communicating parties trust each other. On this basis, the system needs resilience to enhance the robustness in order to better deal with the complex conditions in the real world.

7.2.1 Communication. A grand challenge of collaborative reinforcement learning is how human and AI communicate with each other. Only if the communication is smooth, can they make decisions on the next actions following each other's feedback. Liang *et al.* proposed an implicit human-AI cooperation framework based on Gricean conversational theory [76] to play the game Hanabi. The agent needs to cooperate with the human to win the game. In this framework, the AI tries to understand the implied meaning of human's natural language hints in a dialogue box [43].

Cordona-Rivera and Young proposed an AI Planning-based Gameplay Discourse Generation framework to achieve communication between human players and the game [9]. Pablo and Markus proposed an approach to Human-AI cooperation by planning and plan recognition [10]. Johnson *et al.* proposed a testbed for the joint activity. The unique feature of this testbed is that it can be applied not only in interactive experiments for multiple agents but also in interactive experiments between humans and agents [29]. A series of works were carried out on this testbed to study the cooperation of humans and agents in a team. For example, Matthew *et al.* introduced the relationship between the interdependence and autonomy in a human-AI system [69].

7.2.2 Trust. Based on the established communications, how to make the partners trust each other to complete the task is also crucial. Although the community has not yet proposed a clear definition of trust between humans and agents, it is generally regarded as a psychological state [55]. Johnson *et al.* proposed a Coactive Design framework for human-AI joint activities. In their approach, the authors proposed a cooperation system following the perspective of observability, predictability and directability [28]. Observability means that both pertinent parties can observe each other's state and perceive the environment. Predictability means that both partners can predict the scope of the next actions from the other partner, and these actions are trustworthy.

7.2.3 Resilience. Resilience is another essential feature in human-AI cooperation. On the premise of communication and mutual trust, in complex problems, with possible delays and information noise, how to establish a resilient mechanism to make the system more robust is very important. An effective human-machine cooperation mechanism should be able to diagnose a problem quickly and provide corrective explanations after the problem occurs so that the system can go back to its track resiliently [28]. Zieba *et al.* proposed a mechanism to measure the resilience of human-machine systems, that is, the ability to anticipate, avoid and recover from accidents to a normal state [79]. This is instructive to

design a cooperative system, as it is necessary to consider how the system responds to emergencies and thus recovers quickly.

7.3 Debative Level Cooperation

Debative model means that humans and agents have different opinions on the task, and they debate based on their different knowledge and understandings, to come up with the best solution. There usually are the following requirements for the model. First, humans and agents have a unified goal, and the realisation of the goal is the primary task. The debate without a unified goal is meaningless for both parties. Second, both parties have insights into a problem based on their own cognitive models and solid reason for their decision. Third, both parties can explain their decisions to each other and communicate effectively. Communication and interpretability are the premises of the debate. Fourth, there are clear evaluation criteria to measure the outcome from a debate to ensure an optimal result. Fifth, both parties can learn and adjust their own knowledge after a debate to achieve better results later [55].

As knowledge-based decisions are fragile and controversial, it is necessary to debate the results [32]. In a complex and uncertain environment, a full debate will better demonstrate the advantages and disadvantages of different decisions. Cooperation at the level of debate requires that both humans and AI agents have sufficiently high professionalism in a specific complex field. Reinforcement learning algorithms based on this level are scarcely mentioned in the literature, but we believe that this form of cooperation will track more attention with the advancement in this field.

Geoffrey *et al.* introduced a framework that enables two agents to debate with each other, and finally, a human judge chooses whom to trust [26]. Although this framework has not yet been applied to the debate between humans and agents, it complies with several specifications described above. In their experiment, the two agents tried to persuade human judges to believe their judgments on the MNIST data [13]. First, the goal of the two agents were unified. Second, the two agents have different judgments based on their own algorithmic perceptions. Third, both agents can generate simple explanations to persuade human judges. Fourth, human judges have intuitive knowledge to make clear judgments. This experiment is enlightening for future research, especially in human-agent and multi-agent debate cooperation.

8 FUTURE WORK RECOMMENDATIONS

Reinforcement learning seems to have reached a plateau after experiencing a rapid development. It is difficult to improve the efficiency of AI agents in a complex environment without clear feedback. The research community has proposed some collaborative methods to overcome these obstacles. For example, humans deliver feedback to AI agents through hardware or sensors to improve algorithms efficiency; AI agents provide humans with explanations of decisions to improve the credibility of algorithms. However, research in this area is only at the beginning stage, and there are many open challenges that need to be paid attention to. In the following sections, we recommend several promising future research directions in the field of Collaborative Reinforcement Learning (CRL).

Combining Different Interactive Methods. Develop more natural multi-feedback interactive methods by studying the advantages and disadvantages of different interactive methods. Single interactive methods would have higher requirements from humans and could be inefficient; whereas multi-modal interactive methods would lower interaction barriers and improve efficiency, providing users with a better interactive experience [5].

In the design patterns mentioned above, 'Combining Different Interactive Methods' belongs to 'Augmentative Level Cooperation'. It is the basis for the application of cooperation technology in real-life scenarios; it also is an important

factor to improve user experience. Therefore, researchers could work on more advanced interaction modes based on this design concepts and applied to different scenarios.

User Modelling. It is important to build generic user models to enable the system to accept user feedback robustly. Such models could be used to build human-AI collaboration applications that reduce human fatigue by detecting and predicting human behaviour [6], due to their ability of adapting to interaction channels and feedback types according to the user's preferences. This would require empirical studies to find a way to map between user types and their preferred interaction channels and feedback types.

In the patterns mentioned above, user modelling is one of the most important issues of 'Integrative Level Cooperation'. Only with accurate models could humans and AI agents communicate without barriers. The ability of predicting each other's behaviour could generate trust. Understanding the unexpected situations that may occur for each participant could establish a more flexible relationship and improve the entire system's robustness. Due to the rapid development of AI in recent years, there is still a lot of user modelling work that has not been carried out. Researchers could build AI and user models based on the perspective of 'Integrative Level Cooperation': communication, trust, and resilience.

Safe Interactive RL. Despite the empirical success of reinforcement learning algorithms, we have very little understanding of the way such 'black-box' models work. This means that the system cannot be responsible for their own decisions [18]. Therefore, how to establish a mechanism to protect human safety and avoid unknown discrimination becomes very important. Solving this problem is crucial for the use of interactive reinforcement learning in high-dimensional environments in the real world.

In the patterns mentioned above, safe interactive RL is an important issue for the 'Integrative Level Cooperation' and the 'Debative Level Cooperation'. In terms of Integrative Level Cooperation, how to ensure the safety of humans is one of the key factors for humans to trust AI. In terms of Debative level Cooperation, when the decisions of humans and AI agents are inconsistent, how to protect the interests is not only an engineering issue, but also an ethical issue. This issue requires the joint efforts of multiple disciplines such as law, sociology, and ethics.

Dynamic Mental Models. The cooperative model requires a dynamic mental model from both humans and AI agents, as they constantly observe and learn about each other. It also needs to update strategies immediately during the learning process [15]. With the increase of cooperation experience, it would be useful to inject experience into the new learning process. At this stage, the implicit human prior knowledge needs to be gradually transformed into explicit experience guidelines to future tasks. Therefore, establishing a dynamic mental model may greatly promote the development of human-AI cooperation.

In the patterns mentioned above, dynamic mental models are critical to the three levels, making dynamic adjustments according to different collaboration levels, which has a great effect on reducing power consumption and improving efficiency [56]. This is a huge challenge for researchers. Researchers could design general dynamic mental models according to the taxonomy we provide.

9 CONCLUSIONS

In this paper, we have presented a survey of Collaborative Reinforcement Learning (Collaborative RL, or CRL) to empower the research into human-AI interactions and cooperative designs. We also have provided a *new CRL classification method* (see Table 1) and a *new CRL taxonomy* (see Figure 2) as a systematic modelling tool for selecting and improving new CRL designs. Despite the great progress made in the field of machine learning, there are still many challenges

that need to be addressed before a generic collaborative model is proposed. To conclude, in recent works, many new interactive methods have been introduced, such as the multi-mode that combines voice, gesture, sound and vision. Many cooperative patterns based on different disciplines have also been proposed, such as mental models based on cognitive engineering and psychological modeling. Through this survey, we provide researchers and practitioners with the tools to start improving and creating new designs for CRL methods.

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[2] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *Ai Magazine* 35, 4 (2014), 105–120.

[3] Riku Arakawa, Sosuke Kobayashi, Yuya Unno, Yuta Tsuboi, and Shin-ichi Maeda. 2018. Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback. *arXiv preprint arXiv:1810.11748* (2018).

[4] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robotics and autonomous systems* 57, 5 (2009), 469–483.

[5] Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L Littman. 2019. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257* (2019).

[6] Christian Arzate Cruz and Takeo Igarashi. 2020. A Survey on Interactive Reinforcement Learning: Design Principles and Open Challenges. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1195–1209.

[7] AV Bernstein and EV Burnaev. 2018. Reinforcement learning in computer vision. In *Tenth International Conference on Machine Vision (ICMV 2017)*, Vol. 10696. International Society for Optics and Photonics, 106961S.

[8] Gwern Branwen. 2020. GPT-3 Creative Fiction. (2020).

[9] Rogelio Enrique Cardona-Rivera and Robert Michael Young. 2014. Games as conversation. In *Tenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

[10] Pablo Sauma Chacón and Markus Eger. 2019. Pandemic as a challenge for human-AI cooperation. In *Proceedings of the AIIDE workshop on Experimental AI in Games*.

[11] Andreas Classen, Patrick Heymans, and Pierre-Yves Schobbens. 2008. What's in a feature: A requirements engineering perspective. In *International Conference on Fundamental Approaches to Software Engineering*. Springer, 16–30.

[12] Francisco Cruz, German I Parisi, Johannes Twiefel, and Stefan Wermter. 2016. Multi-modal integration of dynamic audiovisual patterns for an interactive reinforcement learning scenario. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 759–766.

[13] Li Deng. 2012. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine* 29, 6 (2012), 141–142.

[14] Rudolf Ferenc, Arpad Beszedes, Lajos Fulop, and Janos Lele. 2005. Design pattern mining enhanced by machine learning. In *21st IEEE International Conference on Software Maintenance (ICSM'05)*. IEEE, 295–304.

[15] John R Frederiksen, Barbara Y White, and Joshua Gutwill. 1999. Dynamic mental models in learning science: The importance of constructing derivational linkages among models. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 36, 7 (1999), 806–836.

[16] Sandra Clara Gadanho. 2003. Learning behavior-selection by emotions and cognition in a multi-goal robot task. *Journal of Machine Learning Research* 4, Jul (2003), 385–412.

[17] Yang Gao, Jan Peters, Antonios Tsourdos, Shao Zhifei, and Er Meng Joo. 2012. A survey of inverse reinforcement learning techniques. *International Journal of Intelligent Computing and Cybernetics* (2012).

[18] Javier Garcıa and Fernando Fernández. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16, 1 (2015), 1437–1480.

[19] Elizabeth Gibney. 2016. Google AI algorithm masters ancient game of Go. *Nature News* 529, 7587 (2016), 445.

[20] Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. 2016. Affective personalization of a social robot tutor for children's second language skills. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 3951–3957.

[21] Prasoon Goyal, Scott Niekum, and Raymond J Mooney. 2019. Using natural language for reward shaping in reinforcement learning. *arXiv preprint arXiv:1903.02020* (2019).

[22] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in neural information processing systems*. 2625–2633.

[23] Mark K Ho, Michael L Littman, Fiery Cushman, and Joseph L Austerweil. 2015. Teaching with rewards and punishments: Reinforcement or communication?. In *CogSci*.

[24] Erik Hollnagel and David D Woods. 1983. Cognitive systems engineering: New wine in new bottles. *International journal of man-machine studies* 18, 6 (1983), 583–600.

[25] Andreas Holzinger. 2018. From machine learning to explainable AI. In *2018 world symposium on digital intelligence for systems and machines (DISA)*. IEEE, 55–66.

[26] Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899* (2018).

[27] Charles Isbell, Christian R Shelton, Michael Kearns, Satinder Singh, and Peter Stone. 2001. A social reinforcement learning agent. In *Proceedings of the fifth international conference on Autonomous agents*. 377–384.

[28] Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis. 2014. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction* 3, 1 (2014), 43–69.

[29] Matthew Johnson, Catholijn Jonker, Birna Van Riemsdijk, Paul J Feltovich, and Jeffrey M Bradshaw. 2009. Joint activity testbed: Blocks world for teams (BW4T). In *International Workshop on Engineering Societies in the Agents World*. Springer, 254–256.

[30] Artúr István Károly, Róbert Fullér, and Péter Galambos. 2018. Unsupervised clustering for deep learning: A tutorial survey. *Acta Polytechnica Hungarica* 15, 8 (2018), 29–53.

[31] Uri Kartoun, Helman Stern, and Yael Edan. 2010. A human-robot collaborative reinforcement learning algorithm. *Journal of Intelligent & Robotic Systems* 60, 2 (2010), 217–239.

[32] R KLING. 1981. ROUTINE DECISION-MAKING-THE FUTURE OF BUREAUCRACY-INBAR, M.

[33] William Bradley Knox. 2012. Learning from human-generated reward. (2012).

[34] W Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*. 9–16.

[35] Jens Kober, J Andrew Bagnell, and Jan Peters. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* 32, 11 (2013), 1238–1274.

[36] Gregory Kuhlmann, Peter Stone, Raymond Mooney, and Jude Shavlik. 2004. Guiding a reinforcement learner with natural language advice: Initial results in RoboCup soccer. In *The AAAI-2004 workshop on supervisory control of learning and adaptive systems*. San Jose, CA.

[37] Guillaume Lample and Devendra Singh Chaplot. 2017. Playing FPS games with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[38] Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871* (2018).

[39] Pascal Lessel, Maximilian Altmeyer, Lea Verena Schmeer, and Antonio Krüger. 2019. " Enable or Disable Gamification?" Analyzing the Impact of Choice in a Gamified Image Tagging Task. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[40] Guangliang Li, Hamdi Dibeklioğlu, Shimon Whiteson, and Hayley Hung. 2020. Facial feedback for reinforcement learning: a case study and offline analysis using the TAMER framework. *Autonomous Agents and Multi-Agent Systems* 34, 1 (2020), 1–29.

[41] Guangliang Li, Randy Gomez, Keisuke Nakamura, and Bo He. 2019. Human-centered reinforcement learning: a survey. *IEEE Transactions on Human-Machine Systems* 49, 4 (2019), 337–349.

[42] Mao Li, Tim Brys, and Daniel Kudenko. 2018. Introspective Reinforcement Learning and Learning from Demonstration.. In *AAMAS*. 1992–1994.

[43] Claire Liang, Julia Proft, Erik Andersen, and Ross A Knepper. 2019. Implicit communication of actionable information in human-ai teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[44] Guiliang Liu, Oliver Schulte, Wang Zhu, and Qingcan Li. 2018. Toward interpretable deep reinforcement learning with linear model u-trees. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 414–429.

[45] Robert Loftin, Bei Peng, James MacGlashan, Michael L Littman, Matthew E Taylor, Jeff Huang, and David L Roberts. 2016. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous agents and multi-agent systems* 30, 1 (2016), 30–59.

[46] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, David Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive learning from policy-dependent human feedback. *arXiv preprint arXiv:1701.06049* (2017).

[47] Richard Maclin and Jude W Shavlik. 1996. Creating advice-taking reinforcement learners. *Machine Learning* 22, 1-3 (1996), 251–281.

[48] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. Explainable reinforcement learning through a causal lens. *arXiv preprint arXiv:1905.10958* (2019).

[49] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[50] Inhyuk Moon, Myungjoon Lee, Jeicheong Ryu, and Museong Mun. 2003. Intelligent robotic wheelchair with EMG-, gesture-, and voice-based interfaces. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, Vol. 4. IEEE, 3453–3458.

[51] Anis Najar and Mohamed Chetouani. 2020. Reinforcement learning with human advice. A survey. *arXiv preprint arXiv:2005.11016* (2020).

[52] Erika Puiutta and Eric Veith. 2020. Explainable Reinforcement Learning: A Survey. *arXiv preprint arXiv:2005.06247* (2020).

[53] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E McCullough, and Rashid Ansari. 2002. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)* 9, 3 (2002), 171–193.

[54] David Rohde, Stephen Bonner, Travis Dunlop, Flavian Vasile, and Alexandros Karatzoglou. 2018. Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising. *arXiv preprint arXiv:1808.00720* (2018).

[55] Kjeld Schmidt, J Rasmussen, B Brehmer, and J Leplat. 1991. Cooperative work: A conceptual framework. *Distributed decision making: Cognitive models for cooperative work* (1991), 75–110.

[56] Gesina Schwalbe and Martin Schels. 2020. A survey on methods for the safety assurance of machine learning based systems. In *10th European Congress on Embedded Real Time Software and Systems (ERTS 2020)*.

[57] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815* (2017).

[58] Aaron Sloman. 1986. Did Searle attack strong strong or weak strong AI. *Artificial Intelligence and Its Applications, John Wiley and Sons* (1986).

[59] Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *arXiv preprint arXiv:2009.01325* (2020).

[60] Shweta Suran, Vishwajeet Pattanaik, and Dirk Draheim. 2020. Frameworks for Collective Intelligence: A Systematic Literature Review. *ACM Computing Surveys (CSUR)* 53, 1 (2020), 1–36.

[61] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[62] Matthew E Taylor, Halit Bener Suay, and Sonia Chernova. 2011. Integrating reinforcement learning with human demonstrations of varying ability. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*. 617–624.

[63] Gerald Tesauro. 1995. Temporal difference learning and TD-Gammon. *Commun. ACM* 38, 3 (1995), 58–68.

[64] Andrea L Thomaz and Cynthia Breazeal. 2006. Adding guidance to interactive reinforcement learning. In *Proceedings of the Twentieth Conference on Artificial Intelligence (AAAI)*.

[65] Andrea L Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6-7 (2008), 716–737.

[66] Andrea Lockerd Thomaz, Guy Hoffman, and Cynthia Breazeal. 2005. Real-time interactive reinforcement learning for robots. In *AAAI 2005 workshop on human comprehensible machine learning*.

[67] Andrea L Thomaz, Guy Hoffman, and Cynthia Breazeal. 2006. Reinforcement learning with human teachers: Understanding how people want to teach robots. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 352–357.

[68] Karel van den Bosch and Adelbert Bronkhorst. 2018. Human-AI cooperation to benefit military decision making. NATO.

[69] Karel van den Bosch, Tjeerd Schoonderwoerd, Romy Blankendaal, and Mark Neerincx. 2019. Six Challenges for Human-AI Co-learning. In *International Conference on Human-Computer Interaction*. Springer, 572–589.

[70] Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. 2018. Programmatically interpretable reinforcement learning. *arXiv preprint arXiv:1804.02477* (2018).

[71] Richard M Voyles and Pradeep K Khosla. 1999. Gesture-based programming: A preliminary demonstration. In *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C)*, Vol. 1. IEEE, 708–713.

[72] William Yang Wang, Jiwei Li, and Xiaodong He. 2018. Deep reinforcement learning for NLP. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*. 19–21.

[73] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[74] Klaus Weber, Hannes Ritschel, Florian Lingenfelser, and Elisabeth André. 2018. Real-time adaptation of a robotic joke teller based on human social signals. (2018).

[75] Edward C Williams, Nakul Gopalan, Mine Rhee, and Stefanie Tellex. 2018. Learning to parse natural language to grounded reward functions with weak supervision. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1–7.

[76] Deirdre Wilson and Dan Sperber. 1981. On Grice's theory of conversation. *Conversation and discourse* (1981), 155–78.

[77] Mike Wu, Sonali Parbhoo, Michael C Hughes, Ryan Kindle, Leo A Celi, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. 2020. Regional Tree Regularization for Interpretability in Deep Neural Networks.. In *AAAI*. 6413–6421.

[78] George Zarkadakis. 2015. *In our own image: will artificial intelligence save or destroy us?* Random House.

[79] Stéphane Zieba, Philippe Polet, Frédéric Vanderhaegen, and Serge Debernard. 2010. Principles of adjustable autonomy: a framework for resilient human–machine cooperation. *Cognition, Technology & Work* 12, 3 (2010), 193–203.